DOCUMENT RESUME

ED 272 523 TM 860 428

AUTHOR Coffman, William E.
TITLE Taming the Rasch Tiger: Using Item Response Theory in Practical Educational Measurement.
PUB DATE 17 Apr 86
NOTE 8p.; Paper presented at the symposium, "Taming the Rasch Tiger: Using Item Response Theory in Practical Educational Measurement," at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-20, 1986). For other symposium papers, see TM 860 424-428.
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Adaptive Testing; *Computer Assisted Testing; *Item Banks; *Latent Trait Theory; Mathematical Models; Scaling; *Test Construction; *Test Theory
IDENTIFIERS *Rasch Model

ABSTRACT

The symposium, "Taming the Rasch Tiger: Using Item Response Theory in Practical Educational Measurement," was organized to deemphasize the technical complexities of item response theory (IRT) and to show the audience how IRT can be used in practical educational measurement. Four papers from the symposium are summarized and comments are made on the uses of the Rasch model. The following points must be considered when using items from item pools that are scaled using the Rasch model: (1) the question of fit to the model is an empirical question to be established and not a theoretical answer to be assumed; (2) achievement test items will not easily fit the Rasch model unless the item pool is divided into a large number of independent sets, one for each specific objective; (3) the interpretation of the numbers produced by the application of the model is important and not the numbers generated; and (4) when proposing to use the Rasch model in generating test scores, recognize that one should understand the model and its implications in order to make sensible interpretations. (JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED272523

TAMING THE RASCH TIGER: USING ITEM RESPONSE THEORY IN PRACTICAL EDUCATIONAL MEASUREMENT

(Symposium, NCME, Thursday, April 17, 1986 - San Francisco

DISCUSSION - William E. Coffman, University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 428

TAMING THE RASCH TIGER: USING ITEM RESPONSE THEORY IN PRACTICAL EDUCATIONAL MEASUREMENT

(Symposium, NCME, Thursday, April 17, 1986 - San Francisco)

DISCUSSION - William E. Coffman, University of Iowa

In reviewing the papers for this symposium, I was reminded of an experience I had some thirty years ago at a session on test theory at ETS with Dr. Fred Lord. At that time Dr. Lord was already deeply involved in the development of item response theory, but most of the test development at ETS was taking place in the context of classical test theory. I recall that I found myself deeply concerned by a consciousness of how inadequate classical test theory was. The theory assumed that any test consisted of a random sample of a population of test items and that the norms were based on a random sample of a population of individuals to be measured by the test, and it was quite clear, particularly in the case of test items, that we weren't dealing with random samples. I expressed my concern and Lord replied, "Yes, it's true that the model is an over-simplification of what we actually do, but it's really surprising how well the model usually works." As we have gained experience over the years, we have learned a great deal about the possibilities and limitations of the classical model, and competing theories have been developed that promise to overcome some of those limitations.

The symposium today is concerned with one of these competing test theories, one that is also based on an oversimplified model, and again we are told that it really works pretty well, even when the assumptions are not met very well. Our speakers tell us that the model overcomes some of the inadequacies of the classical model we have used for so long. The purpose of the symposium, according to Dean Forbes, was to introduce test directors and other persons involved in the operation of public school testing programs to the basic concepts involved in the so-called Rasch model. It was felt that many potential users are intimidated by Rasch and consider it too esoteric and abstract either for use or understanding. As a result, Forbes was given the charge to organize a symposium that would de-emphasize the technical complexities and let the audience see how I.R.T. can be used in practical educational measurement.

Let's try to summarize what the speakers have been telling us. Forbes has told us that the Rasch model assumes that, for a test that fits the model, a test score is the result of only two parameters, the difficulty of the items in the test and the ability of the test taker, and that both ability and difficulty are independent of the sample of individuals who may have taken the test and of the sample of items that are in a particular test. Also, that one the item difficulties have been determined, one can estimate the ability of any test taker very efficiently by simply administering items that are close in difficulty to the ability of the test taker. Furthermore, since the ability and difficulty indices are sample free, once a bank of scaled items is created, the task of building a test is very simple: just select items of appropriate difficulty in relation to the ability of the examinees. Finally, since computer programs are available to do the job of providing the estimates, one doesn't need to understand the complexities of the theory in order to make use of the applications.

Hunt goes a step further. Given that the item bank fits the model, one is free to select from the bank only those items that are judged to be relevant to the local curriculum. Furthermore, "cross linking with national norms may be developed for a broader comparison." She notes the problem that current item banks deal only with basic skills areas and that there is need for item banks in the content fields such as science and social studies. Neither Forbes nor Hunt provide any data to support their arguments, but Forbes does cite references to the work of others.

Mershon reports on the experience of a school system in the development of criterion referenced tests and how the local item pool was expanded by relating it to more comprehensive item pools so that local items could use the scales that underly the larger pools and so that items from the pools might be used for local tests. The paper introduces a note of caution: that for any particular goal, a test score should be based on a minimum of 20 items if one wishes to obtain acceptable reliability of the score. He does not, however, examine the question of how items that measure a variety of different goals can be scaled together in a common pool.

In the final paper, Thomas raises a number of questions about a tendency to accept claims for the model without question. He reports that item difficulty indices are not typically as stable as the Rasch theory would have us believe. Item difficulty indices do turn out to be different when they

are calibrated on samples of different ability. Also, they differ from time to time, even for comparable samples of test takers. And since the item statistics for a pool can "drift", it's probably a good idea to have as an additional reference point data from a nationally normed test. Then too, since locally developed test items may vary in quality, it is desirable to be sure that they do fit the model well enough so that one can have some confidence in the scaling. Those items that do not fit the model need to be discarded. However, he continues, "if we repeatedly delete items from tests which are not meeting mathematical model expectations, over the course of time we may be producing a test which is free of curricular effects. Perhaps following this approach in an attempt to produce a 'tight' metric serves the purpose of retaining items which are in fact not impacted by those modifications to curricula which we continually strive to make."

One sentence in Thomas' paper warrants special attention. He writes, "the difficulty of some items remained stable across different populations (italics mine) of students while some item difficulties changed." Here Thomas has put his finger on a critical issue in evaluating the appropriateness of basing testing practices on the Rasch model. The theory is that when the data fit the model, item difficulty indices and person ability estimates are sample free, not population free. This means that one who proposes to use the Rasch model needs to think very clearly about whether or not the individuals for whom the test is intended belong to the population on which the items were scaled -- or whether the particular set of items in the test are still measuring the same ability for these individuals that they were measuring for the individuals in the scaling group.

Those who are most enthusiastic about the use of the Rasch model assure us that the model is robust and that one doesn't need to be concerned about the fact that the model is an oversimplification of the situation in the real world of educational tests. The computer can take care of the mathematical details. All you have to do is trust the programs to produce numbers that tell you what you want to know. But are there dangers one needs to be aware of?

Many years before the Rasch model was proposed, Ledyard Tucker was already studying the appropriateness of item characteristics curves as describers of test questions and was conducting research on the scalability of test items. He reasoned that English vocabulary items were likely to be scalable if any were, but he was also aware that he ought to have a clear idea of whether or

not they would behave in the same way for different subsets of the student population. He therefore carried out the scaling on four different groups: (1) high socioeconomic northeasterners, (2) high socioeconomic southeasterners, (3) low socioeconomic northeasterners, and (4) low socioeconomic southeasterners. He concluded that most vocabulary items would scale, that is, that they had the same relative difficulty from group to group. For most of the items, the relative difficulty was 1, 2, 3, 4 as had been predicted. But there were exceptions. For example, the relative difficulty of the word "scorpion" was 2,1,3,4. Scorpions are more familiar to southeasterners, but the low socioeconomic southeasterners call them "sting lizards." As any student of the Rasch model knows, items that don't fit the model are to be kept out of the pool. It is questionable, however, whether the experimental procedures used to determine fit are as appropriate as those used by Tucker many years ago.

One might raise the question, for example, of whether or not the same scale values would be obtained if one were to take 10% of the vocabulary items from a pool, drill the students in a particular school on these vocabulary words only for three weeks, and then scale the items again using only those students. There is also the question of the effects of the passage of time on the difficuly of certain test items. I recall that we found for items in the College Board Social Studies pool dealing with presidential elections the p-values varied in four-year cycles corresponding to the years of presidential elections. I believe a similar pattern would be found with Rasch difficulty indices.

These problems are well understood by those who deal with test theory. For example, I found this statement on page 66 of Best Test Design by Wright and Stone:

> Even if the measurement model tends to fit a particular application, we cannot predict in advance how well new items (or even old ones) will continue to work in every situation in which they might be applied, not can we know in advance how all persons will always respond. Therefore, if we are serious in our attempt to measure, we must examine every application to see how well each set of responses corresponds to our model expectations. We must evaluate not only the plausibility of the sample persons' responses, but also the plausibility of each person's responses to the set of items in his test. To do this we must examine the response of each person to each item to determine whether it is consistent with the general pattern of responses observed.

I see little evidence that those who are enthusiastic about applications of the Rasch model are following this recommendation.

To those of you who are considering the use of items from item pools that are scaled using the Rasch model, let me urge these cautions:

1) When the individuals being measured are a part of the population that manifests the ability defined by the pool of items that defines the ability, both ability and difficulty are sample free; but the question of fit to the model is an empirical question to be established and not a theoretical answer to be assumed.

2) Often, in applying the Rasch Model to develop a comprehensive scale for an item pool, the assumption is made that there is an approximately normal distribution of ability in the population and an approximately normal distribution of difficulty for the item pool. Normal distributions are most likely to occur as a result of the operation of a large variety of elements over a long period of time--as in the development of general language ability. Any limited number of systematic elements, as in the application of a carefully organized teaching effort, is likely to produce departures from the model. In fact, one might argue that the purpose of a school curriculum is to produce learning that is non-random and therefore learning that is distributed in other than a normal pattern. I suspect that the more we learn about how to teach, the less likely it is that achievement test items will fit the Rasch model, that is, unless we divide the item pool into a large number of independent sets, one for each specific objective.

3) Recall that the Rasch model is based on the same basic information as the classical model, that is, the responses of individuals to test items; and there is nothing magical about the numbers that are produced by application of the model. You still have to think about what interpretations are legtimate for the numbers generated by your application of the model. Major advantages of the model are that it permits one to focus measurement at particular regions of the ability scale, that it is possible to assess the efficiency of the measurement at different points on the scale, and that an analysis of responses can tell you when things are going wrong. Unfortunately, in order to know what to pay attention to in the analysis, you have to understand the model, not simply to let the computer produce numbers to be interpreted automatically.

In this context, it is interesting to note that a Rasch analysis has been used to identify items that are biased against certain minority groups. Notice, however, that when such items are identified, it is not always appropriate simply to drop the discrepant items from the test, assuming that the scaling of the

remaining items is correct. It depends on how much the items that are dropped contributed to the original scaling. Of course, there is always the possibility that if you rescale the remaining items, the scale value won't change very much. Certainly this is likely to be the case if there are only a few items identified as biased in a large pool of items. If however, one is choosing only a limited number of items from a pool that are identified as appropriate because they are judged to measure those objectives taught in a particular curriculum, then those items may actually be defining a different ability from that represented by the original pool of items.

4) Finally, if you propose to make use of the Rasch model in generating your test scores, recognize that the better you understand the model and its implications, the more likely you are to make sensible interpretations of the test scores you generate. Don't be mesmerized by the sales pitches of those who are marketing item pools or computer programs. Insist on understanding what it is you are buying before you close the deal.